DOCUMENT RESUME

ED 124 574 TM 005 327

AUTHOR Barta, Maryann B.; And Others
TITLE Some Problems in Interpreting Criterion Referenced Test Results in a Program Evaluation.
PUB DATE [Apr 76]
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Achievement Tests; *Comparative Analysis; *Criterion Referenced Tests; Intermediate Grades; Norm Referenced Tests; *Program Evaluation; *Standardized Tests; Student Testing; Testing Problems; *Test Interpretation; *Test Results

ABSTRACT

This paper delineates several problems which arise when criterion-referenced test results are used to evaluate the effects of a specific educational treatment. Focus is on these topics: (1) the aggregation of individual students' data on objectives, (2) the sensitivity of the instrument to instruction, (3) the interpretation of criterion-referenced group data, and (4) comparison of crition-referenced test results and standardized test results. The results show that the methods used in reporting individual objective data affect the outcomes. In addition, comparisons of gains show that the criterion-referenced tests are not necessarily more sensitive to student growth. (Author/RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

26.10

ED124574

SOME PROBLEMS IN INTERPRETING CRITERION REFERENCED TEST RESULTS IN A PROGRAM EVALUATION

by

Maryann B. Barta
Unhai R. Ahn
Joseph F. Gastright
Department of Research & Development
Cincinnati Public Schools

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Presented at Annual Meeting of the
American Educational Research Association
San Francisco, California

April, 1976

TM005 327

SOME PROBLEMS IN INTERPRETING CRITERION REFERENCED TEST RESULTS IN A PROGRAM EVALUATION

Maryann B. Barta, Unhai R. Ahn, and Joseph F. Gastright, Cincinnati Public Schools.

Purpose

The purpose of this paper is to delineate several problems which arise when criterion referenced test results are used to evaluate the effects of a specific educational treatment. Specifically, the paper deals with: (1) alternative methods of aggregating individual student and group data on objectives, (2) the sensitivity of the instrument to program outcomes, and (3) the comparisons of criterion referenced test data and standardized achievement test data.

Background

During the past decade, there has been extensive discussion of the merits of criterion referenced testing as an alternative to norm referenced tests (Popham & Husek, 1969; Hambleton & Novick, 1972; and Gronlund, 1973). While criterion referenced tests have been defined in a multitude of ways, an underlying thread among all of these definitions is the assumption that criterion referenced tests are deliberately constructed so as to yield measurements that are directly interpretable in terms of specified performance standards (Glaser & Novick, 1971). In spite of this assumption of direct interpretability, very little clear direction is given in the literature of specific ways in which criterion referenced test results have been used practicably

to evaluate either student progress or program outcomes. Popham & Husek (1969) recommend that a number of schemes to report the group's performance be employed in order to permit more enlightened interpretations; for example, the number of individuals who achieve the criterion, traditional descriptive statistics such as the mean and standard deviation, and an average "percentage correct." Knipe & Krahmer (1973) present "student by objective grids as unsophisticated ways of detecting different learning patterns." Gronlund (1973) recommends that criterion referenced test results be interpreted cautiously.

Empirical examples of criterion referenced test results reported in the literature (Hsu, 1971; Roudabush, 1973; and Roudabush & Green, 1971) have focused on the improvement of criterion referenced test items, rather than the use of the data for instructional decisions. The extensive discussion of criterion referenced test reliability and errors of measurement (Millman et al, 1975) suggests that criterion referenced test results may be far from directly interpretable.

Statement of Problems

The literature has contained many articles about the controversy between criterion referenced tests (CRT) and norm referenced tests. Most of these articles were based on the conceptual and theoretical differences between these tests. Few of the articles made objective comparisons based on empirical data.

Both norm referenced and criterion referenced tests are designed to make decisions about individuals or programs. The decision may be one of selection or one of improvement. In the case of norm referenced tests, the decisions are made in reference to the performance of

normative groups of individuals or programs placed in the same decision situation. In the case of criterion referenced tests, the decision is critically related to a comparison of the individual's performance with an arbitrarily established standard of performance or criterion level. This latter point becomes important when the decisions are made on test items which are not obviously norm-distinctive or criterion-distinctive. The items on two types of tests are, in fact, more often interchangeable than not.

Unlike other papers on criterion referenced tests and norm referenced tests, this paper is based on empirical data collected concurrently with both a Criterion Referenced Test and a Norm Referenced Test. Some of the questions that the study will address are:

1. If one reports and aggregates criterion referenced test data in different ways, would the results be consistent?
2. Is the criterion referenced test sensitive to the changes that occur in students?
3. Are the estimates of the program effects based on criterion referenced test results and standardized test results comparable?

Methods

Data were collected on a group of 182 fourth, fifth, and sixth grade students located in two elementary schools within the Cincinnati Public School System. These students were selected for this study because of their involvement in a commercially prepared reading comprehension and verbal skills curriculum.

The curriculum is an individualized, self-paced program for the development of reading skills. Each student proceeds at his own pace with a prescribed set of learning materials and activities provided in the reading learning center.

Each student participating in the program was tested with a commercially prepared Criterion Referenced Test in November, 1974, and again in May, 1975. The Criterion Referenced Test was designed by a reputable educational testing firm to assess the objectives of the curriculum. Each student was also tested with an appropriate level of the reading subtest of the Metropolitan Achievement Test in October, 1974, and April, 1975.

The level of the Criterion Referenced Test that was given to the students was determined by their score on a short screening test. There were three levels (I, II, and III) of the Criterion Referenced Test. It was possible for a fourth grade student to take the highest level (III) of the Criterion Referenced Test, and it was just as possible for a sixth grade student to take the lowest level test (I). Data on the criterion referenced test and the standardized achievement test were analyzed separately according to these groups.

Each test level included different objectives. Although there was an overlap of objectives at each level, the test items at each level measuring the objectives were different. Table 1 shows the objectives included at each level and also the number of items measuring each objective at each level.

Mastery of an objective was determined for students who had 75 percent or more of the items correct for the objective. Student progress through the curriculum was determined by the same criterion. Table 2 indicates the rules used in determining the mastery for each objective.

Table 1. Objectives at Each Level of the Criterion Referenced Test and the Number of Items Measuring Each Objective.

Level I	Level II	Level III
Objective (# of items)	Objective (# of Items)	Objective (# of Items)
Letter Recognition (2)	Sentence Comprehension (2)	Contextual Cues (1)
Initial, Final Sounds (5)	Contextual Cues (3)	Main Theme (5)
Vowel Sounds (4)	Main Theme (5)	Specific Detail (5)
Consonant Sounds (6)	Specific Detail (5)	Sequence (3)
Word Endings (3)	Sequence (4)	Drawing Inferences (3)
Other (6)	Drawing Inferences (5)	Author's Intent, Viewpoint (3)
Sentence Comprehension (2)	Author's Intent, Viewpoint (4)	Word Meanings (4)
Main Theme (3)	Word Meanings (5)	Special Usage (3)
Specific Detail (3)	Special Usage (4)	Follow Directions (3)
Sequence (3)	Follow Directions (3)	Interpret Charts, Graphs (3)
Drawing Inferences (2)	Understanding Structure (1)	Understanding Structure (3)
Author's Intent, Viewpoint (3)		Use Content Classifiers (3)
Word Meanings (3)		Paragraph Meaning (4)
Special Usage (2)		

Table 2. Rules for Determining Mastery.

Number of Items Testing an Objective	1	2	3	4	5	6
Items Required for Mastery	1	2	3	3 or 4	4 or 5	5 or 6

There are standard ways of reporting and aggregating standardized achievement test scores for individuals as well as for groups. In this study, the gains on the standardized achievement test were obtained by subtracting the pretest standard score from the posttest standard score for each child. The standard score gains were averaged within levels of the criterion referenced test groups.

However, there is no standard rule for presenting criterion referenced test results. Several alternative methods of analyzing criterion referenced test data are possible in terms of the group's status at the beginning and the end of instruction or gains made during that period. The status of pupils on either the pretest or posttest could be displayed as either the percentage of items correct or the number of objectives mastered by each child. Gain could be calculated either as an increase in the number of items answered correctly or in terms of change in the number of objectives mastered. Figure 1 describes the type of raw data that can be aggregated to produce objective-based gain scores.

Figure 1. Pre and Post Changes on the Criterion Referenced Test.

		Posttest +	Posttest −
Pretest	+	A	C
	−	B	D

\+ Objective mastered

\- Objective not mastered

Cell A are those students who maintained mastery.
Cell B are those students who recently mastered.
Cell C are those students who lost mastery.
Cell D are those students who never mastered.
Cells A + B + C + D = T, the entire student population.

In this study, gain scores for individuals on the criterion referenced test were calculated in two ways: first, a simple raw item gain between the pretests and posttests; and second, as net gain in objectives mastered (B - C) by each student between the pretests and posttests. Each of these gains was correlated with each other and with the gain in standard scores on the standardized achievement test.

Gain scores for groups of individuals were calculated in four ways:

1. Gross Mastery in Total: the percentage of students who achieved mastery on the posttest $\left\langle \frac{A + B}{T} \right\rangle$

2. Gross Gain in Non-Masters: the number of students gaining mastery as a percentage of the non-mastery group on the pretest $\left\langle \frac{B}{B + D} \right\rangle$

3. Gain in Total: the number of students gaining mastery as a percentage of the total group $\left\langle \frac{B}{T} \right\rangle$

4. Net Gain in Total: the number of students gaining mastery minus the number losing mastery as a percentage of the total group $\left\langle \frac{B - C}{T} \right\rangle$

Obviously, the data could be reported for each objective or aggregated over all of the objectives at each level. The emphasis of this report is on using the data for program evaluation. Therefore, the data is aggregated over individuals and objectives for presentation. The average values of A, B, C, and D in terms of objectives for students in each group were calculated. The data were also presented as percentage of mastery according to the four data presentation methods defined above.

Results

The matrix of students-by-objectives on a criterion referenced test reported on a mastery or non-mastery basis clearly has a diagnostic instructional value. In an instructional situation in which all of the students begin with non-mastery, the proportion of students gaining mastery across the instructional period becomes a measure of the impact of the program. In most group situations, however, the assumption of non-mastery prior to instruction cannot be made. In relatively homogeneous groupings of students, as might be obtained by using a screening device, some students will achieve mastery on initial testing, while other students will not. The proportion of non-masters (D and B) who subsequently achieve mastery (B) provides a relatively optimistic estimate of the effects of the program. In these cases, the number of individuals gaining mastery is more revealing than the actual percentage of pupils gaining mastery, since the percentage values can be inflated if most of the students achieved mastery prior to the instructional sequence. The percentage of the total group mastering an objective across the instructional period provides a more balanced measure of program impact. However, it may look artificially low for those objectives mastered by high percentages of students on the pre-measure. All of these measures of gain in mastery fail to describe the impact of instruction in a course where significant loss in mastery (C) occurs in those objectives assessed as mastered on the pre-measure. Table 3 displays the percentages of mastery under the above conditions with objectives accumulated over all students at the three levels.

Table 3. Percentage of Objectives by Group, by Alternative Methods.

Alternative Methods for Criterion Referenced Test Data Analysis		Criterion Referenced Test, Levels I	II	III
1) $\frac{A + B}{T}$	Gross Gain in Total	43%	45%	37%
2) $\frac{B}{D + B}$	Gross Gain in Non-Masters	35	34	28
3) $\frac{B}{T}$	New Gain in Total	27	20	20
4) $\frac{B - C}{T}$	Net Gain in Total	20	4	10

In all cases, a significant proportion of the total number of objectives remained unmastered on the posttest. Overall, the students at each level mastered almost one-third of the objectives which were assessed as non-mastered on the pretest (B/D + B). As might be expected, the level of total mastery found a smaller proportion of the objectives mastered (B/T). When interest is focused on the net gain in mastery (B - C/T), the proportional impact of the program becomes less impressive.

Clearly, the method of reporting mastery has an effect on the interpretation of these results. Methods which concentrate on the impact on non-masters can clearly exaggerate the cumulative impact on refined learning (B/D + B vs. B - C/T).

The sum data can be expressed as an average value for all students on the quantities A, B, C, D, and Total used to calculate the percentages in Table 4.

Table 4. Changes in Average Criterion Referenced Test Objectives, by Level.

Level	Number of Students	Total # of Objectives	Maintained Mastery A	Recently Mastered B	Lost Mastery C	Never Mastered D
I	70	14	2.3	3.8	.9	6.9
II	100	11	2.7	2.2	1.7	4.4
III	12	13	2.1	2.6	1.3	6.9

The large increases in the average number of objectives lost across instruction (C) clearly affect the interpretation of the results as measures of program impact.

The most commonly mentioned advantages of the criterion referenced testing are their diagnostic usefulness in targeting instruction to specific homogeneous objectives and their sensitivity as measures of the effects of the program on these targeted objectives. These advantages were maximized in the curriculum being assessed in the present study. Progress through the pre-programmed curriculum was based on successful attainment of mastery on items and criteria levels which coincide with the items and criteria levels utilized in both the pretest and posttest.

Table 5 gives the pre, post, and gain scores on the criterion referenced test in both items and objectives at each level. These gains correspond to the net gains (B - C) outlined in Table 4. By ordinary measurement standards, the criterion referenced test item gains are, at best, modest. If these gains represent a more sensitive assessment of the true program impact, then the weight of evidence borne by individual items is very high.

Table 5. Pretest, Posttest, and Gain Scores by Criterion Referenced Test Items and Objectives.

Level	N	Pretest		Posttest		Gain	
		Items	Objective	Items	Objective	Items	Objective
I	70	26.6	3.2	30.8	6.1	4.2	2.9
II	100	24.8	4.4	26.0	4.9	1.3	.5
III	12	23.9	3.4	25.3	4.7	1.3	1.3

The final question addressed in the study is whether the results of criterion referenced testing give different estimates of impact than would have been obtained from standardized test results.

Table 6 describes the gain scores by criterion referenced test items, criterion referenced test objectives, standardized test standard scores, and grade equivalents.

Table 6. Mean Gain Scores for CRT and Standardized Test by Group.

Group	N	CRT Objective Gain	Standardized Test	
			Standard Score Gain	Grade Equivalent Gain
I	70	2.9	3.0	.2 years
II	100	.5	5.0	.5 years
III	12	1.3	10.0	1.4 years

The comparison of net gains is quite different across the three levels. The criterion referenced test results would suggest that the program was most successful with the lower level students, second best with the highest level, and worst with the middle level students. Whether the CRT gains are

positive or negative must be determined in relation to some standard that presently is not available. The standardized test results indicate poor gains in reading comprehension for the lowest group; predictable, but not outstanding gains for the middle group; and quite exceptional gains for the admittedly smaller highest group. On the surface, then, the criterion referenced tests do not give the same estimates of program effectiveness as would have been obtained from standardized tests.

The gain scores on the Criterion Referenced Tests by item and objective were correlated with the gains in standard scores on the Standardized Test within each group (Table 7).

Table 7. Intercorrelations Between Gain Scores by Group.

Variable	Group I			Group II			Group III		
	1	2	3	1	2	3	1	2	3
1. Gain in Standard Score	1.00			1.00			1.00		
2. Gain in CRT Items	.00	1.00		.14	1.00		.15	1.00	
3. Gain in CRT Objectives	.14	.84	1.00	.17	.72	1.00	.23	.42	1.00

The results suggest that the gains on the Standardized Test are unrelated to the gains in either items or objectives on the Criterion Referenced Test. The gains in items and objectives on the Criterion Referenced Test were rather strongly correlated in two of the three groups.

Discussion

Admittedly, the situation under which the present data were collected deviates in many respects from an experimental study. No control was possible on the amount of the commercial curricula covered by each student. The students participating in the program proceeded at an individualized pace through the material without regard to external grade level standards. Further, the decision to use the standardized reading comprehension scores for comparison was purely based on the availability of data.

The resulting data are by all standards open to alternative explanations. The program itself may not have been optimally implemented, or implemented in similar fashions in the two sites. The focus of the study is on the concurrent assessment of the impact of the program with two "types" of instruments: criterion referenced test and standardized achievement test.

It is clear that the manner in which criterion referenced test results are aggregated to measure program impact can effect the relative interpretation of the results. Concentration on posttest scores of non-masters can have two-fold pernicious effect on the use of criterion referenced test results. First, this form of reporting tends to exaggerate the estimates of program effectiveness. It takes advantage of a form of a regression effect to the extent that non-masters can only get better when assessed by items with questionable "reliability" to assess objectives against a relatively arbitrary criterion.

Any valid measure of objective-based gain should include reassessment of "mastered objectives" and calculation of a net gain in mastery. The

significant amount of "lost mastery" documented in the present paper dictates reassessment of "presumed mastery."

Another effect of "presuming mastery" is more directed at the diagnostic use of criterion referenced test results. A student who achieves mastery of an objective on the basis of five items may be eliminated from further instruction or reinforcement of the skill involved. If the assessed mastery status was incorrectly made, then the number of students who subsequently "lost mastery" includes a significant number of misidentified students. It is also possible that "mastery" in such a situation is dependent upon continual use of the skill. The assumption that important learning is a one-time event does not seem justified on the basis of existing learning theories.

The present results do not suggest that criterion referenced tests give the same evaluation results as standardized tests. The gains on the criterion referenced test were greatest at the lowest functional level; the gains on the standardized test were greatest at the upper functional level. One could hypothesize on this rather weak evidence that criterion referenced tests are more sensitive to gain in lower level skills, while standardized tests are more sensitive to higher ones. The hypothesis deserves testing in other situations where concurrent data on standardized tests and criterion referenced tests are available. It seems likely that fundamental reading skills are more consistent with a mastery learning model than are more complex behaviors.

The results show that the effects of an instructional program will not always be equally assessed by criterion referenced tests and standardized tests. It has not been conclusively proven that criterion referenced tests will show gain where standardized tests do not. Those

practitioners who turn to criterion referenced tests as a guaranteed measure of "more positive results" will be disappointed occasionally. The contention that learning outcomes are "adequately measured" by comparison of performance on some limited number of test items with some essentially baseless criterion level seems at least as capricious as the basis on which the same decisions are made with standardized achievement tests.